Reference rot in scholarly statement: threat and remedy

Based on a breakout session presented at the 38th UKSG Annual Conference, Glasgow, March 2015

As the scholarly communication system evolves to become natively web based, citations now commonly include hyperlinks to content that is issued on the web. The content at the end of those hyperlinks is subject to what has been termed 'reference rot': a link may break or the content at the end of the link may no longer represent what was first noted as significant. Reference rot threatens both the usability of what is published and the long-term integrity of the scholarly record. The aim of the Hiberlink project has been to focus on this problem and then to compile and analyse a large corpus of full-text publications in order to quantify the extent of reference rot. The results are now out, and the task has shifted to alerting publishers and libraries on what to do in order to ensure that published web-based references do not rot over time. This has implications for the integrity of the scholarly record and for authors of that record. Fortunately, the Hiberlink project has progressed further than originally envisaged and has recommended remedies aimed at alleviating reference rot.

Introduction

What is now cited as a reference in scholarly statement is increasingly 'on the web'. That could mean reference to content stated in formal academic publication, such as an article in an e-journal, for which there is the important infrastructure represented by CrossRef. However, increasingly, what is referenced as significant in the scholarly statement is to be found on the 'web at large' (the 'wild web'). That content on the web lacks the fixity formerly associated with academic publication. What is found on the web at large at any one time is liable to have changed or to have simply disappeared without a trace when time has elapsed and the scholarly statement is published and read. The problem of information decay for citations to content on the internet has been noted by many others in the past, and certainly since the emergence of the web. The term 'reference rot' was coined in the Hiberlink project[1], which set out to study the problem in depth and on a large scale.

What is reported here complements previously published research papers[2, 3] and follows on from workshops delivered at two recent conferences: Electronic Theses and Dissertations 2014 (ETD2014)[4] and the UKSG 38th Annual Conference and Exhibition (UKSG2015).[5] The Hiberlink project was funded by the Andrew W Mellon Foundation and brought together the Los Alamos National Laboratory Research Library and the University of Edinburgh (EDINA and the Language Technology Group of the School of Informatics).

Our purpose here is to provide an explanation of reference rot and to outline and propose a number of remedies that could be taken by publishers when working with authors. University librarians might also wish to take note as the effect of reference rot has even more severe consequences for doctoral dissertations, which are both longer in length than most articles and are prepared over a longer period of time.

Reference rot has two components:

- *link rot*, signified by the familiar and unhelpful '404 – Page Not Found' error message
- *content drift*, either obvious where the link points to something entirely different, or insidious when the content at the end of the link is dynamic (e.g. a news website) and has changed.

PETER BURNHILL

Director
EDINA

MURIEL MEWISSEN

Project Manager
EDINA

RICHARD WINCEWICZ

Software Engineer
EDINA

Understanding the extent of the threat

Web resources have become an integral part of scholarly publication. Citations to software, data sets, websites, ontologies, presentations, blogs, videos, etc. are increasingly common. Reference rot therefore threatens the usability of what is published and, over the long term, the integrity of the scholarly record. This lack of quality should be of concern to publishers and their customers as well as to research libraries and their constituents.

The Hiberlink project assembled a large corpus of data of more than 3.5 million scholarly articles from three different sources: arXiv, Elsevier and PubMed Central. A second corpus was built with 6,400 e-theses downloaded from the repositories of five universities. A workflow was developed to extract the uniform resource identifiers (URIs) from the references made in these articles[6] and then, having excluded URIs that pointed at online journal articles, to assess whether those URIs that pointed to content on the web at large were still active (or had a broken link despite repeated follow-through), and whether the URIs had content that could be found in a web archive in and around the time of publication.[7] Several online archives were used including Internet Archive and archive.is.

References to online journal articles were not considered in this project because the preservation and long-term access of these need to be addressed differently, through arrangements for the archival ingest of the stream of content from serials by digital archiving agencies such as CLOCKSS, Portico and LOCKSS, with various degrees of success, as monitored by the Keepers Registry. We acknowledge that work remains to be done in this area, as noted by David Rosenthal in a blog post on the evanescent web[8] and in work done by the team at the Keepers Registry.[9,10]

One in five articles suffer from reference rot

The results of the Hiberlink investigation[11] are simple and stark. They show that:

- reference to web resources has increased dramatically over the last 15 years
- the probability of references rotting increases with time: the longer the elapsed time since the web content was noted and the article published, the more likely that its references are now rotten
- Of science, technology and medicine (STM) articles published in recent years (2009 to 2012), one in five (20%) suffer from reference rot. These contain at least one reference to a web resource that is rotten, where no archived version of the referenced resource exists within the 14-day time interval of the date of publication of the article.

'one in five STM articles (20%) suffer from reference rot'

Similar figures have been reported by researchers at Harvard Law School on their work on reference rot in legal citations.[12]

As noted, opportunity was taken to examine the much longer form e-theses and dissertations that are prepared for the award of a doctorate. The results of that examination were reported at the ETD2014 conference, a presentation which can be accessed online[13], but for which the study is yet to be summarized and published formally. The manner in which the doctoral thesis/dissertation is a significant part of the scholarly record may vary, but there is little doubt about the importance of the e-thesis for its author and for the university. Preparation takes place over an extended period, giving greater opportunity for reference rot to occur.

In the aforementioned examination, reference rot was assessed in a corpus of 6,400 e-theses, published between 2003 and 2010, downloaded from the institutional repositories of five US institutions. The URIs to web resources were extracted. Of the 46,000 that pointed outward to the web at large, a third were subject to link rot and no longer available on the live web. Using Memento[14], it was discovered that there were archived copies of that content in web archives for no more than half of these. That happy coincidence meant that about half of those might have avoided reference rot. However, what the study also exposed was

that 34% of content currently live on the web was not archived and, therefore, was at risk of being lost. Moreover, 18% of those references should be deemed lost forever, being neither live on the web nor with any evidence of incidental archiving, as shown in Table 1.

	On live web	Not on live web	Total
Archived	29.3%	18.3%	47.6%
	(Safe)	(Preserved)	
Not archived	34%	18.4%	52.4%
	(At risk)	(Lost)	
Total	**63.3%**	**36.7%**	**100%**

Table 1. This table shows the effect of reference rot in a corpus of 6,400 e-theses defended between 2003 and 2010 in five US institutions: Florida State University, University of Notre Dame, Penn State University, Virginia Tech and Worcester Polytechnic Institute. There were 45,982 URIs extracted and tested for link rot and archive status. This shows that 29.3% of URIs are considered safe in term of long-term access, with content available on the live web and in web archives. Of the remainder, 34% of URIs are at risk: although they are live on the web, they are not archived; 18.3% of URIs are preserved but no longer available on the live web; and, finally, 18.4% of URIs are lost for good because they are not on the live web or the web archives.

The empirical evidence of the threat of reference rot in various forms of scholarly statement, and therefore the scholarly record, is overwhelming. The longer the time lapse, the greater is the probability that the referenced content will no longer be at the end of the cited URI. There is less than a 50:50 chance that the cited content will have been archived by routine, incidental web archiving.

'There is less than a 50:50 chance that the cited content will have been archived by routine, incidental web archiving'

What then can be done to prevent reference rot from happening or to stop the rot for references in that scholarly statement which has already been published or otherwise made available?

Remedy

The Hiberlink project identified three basic workflows in which there might be opportunity for productive intervention for the actors involved:

- *the preparation stage:* for the author, when note-taking and writing prior to submission
- *the submission stage:* for processes within the publishing organization, typically via an editor working with the author, overseeing review and issue
- *the post-publication stage:* for use and the record, typically the access platform, the archiving organization and the library, in receipt of published/issued work.

Hiberlink investigated what software processes and information infrastructure, used in conjunction with extant tools, would assist the respective actors to mitigate the risk that is currently being perpetuated.

What to do and when to act

This following advice was set in an earlier enquiry when discussing possible solutions with colleagues in Edinburgh:

'As ever more information resources become available on the Web, the need for effective preservation solutions continues to grow. The case therefore also grows for academic authors – students and researchers – to acquire the habit of referencing stable, reliable copies of them, rather than copies 'in the wild', which can easily mutate, or disappear without trace.' (Davis, 2010)[15]

Stage 1: preparation

The best time to avoid the rot is to act when the reference is first consulted and regarded as significant. This means creating snapshots (mementos) of web pages either at the authoring

stage or at the submission/review stage. Often, reference rot will have already begun by the time an article has been published.

'reference rot will have already begun by the time an article has been published'

It is unrealistic to expect authors to do this unaided. Instead, the plan is to build the means to proactively archive temporal references into, or enabled from, existing tools that researchers and research students already use. Such tools include reference management software like Zotero, EndNote and Mendeley. Outreach to the providers of these tools has already begun.

Zotero plugin demonstrator

Zotero[16] was the first choice for software engineering effort at EDINA because it is free and open source. This made it easy to obtain and extend the code in order to embed the workflow that would manage the interaction with one or more web archiving organizations. The goal was to assist the author in creating an archival copy of any web page regarded as significant, with minimum burden or disruption. This has been achieved by processing the archiving request in the background, allowing the user to continue to work, and prepare a more sustainable reference link. Once the web archiving organization creates the archival copy, its URI is passed back to the reference manager which then adds the URI of the archival copy, and the date and time at which the archival copy was created, to the original URI in the citation record. This information is then available for the author to export and use as part of the citation.

Stage 2: submission

The second best time to act is when the author has completed the manuscript and is in the process of entering this into some type of submission system. The role of the editor and the 'copy-editing' aspect of interaction between author and editor may be crucial. Incorporating an archival step into a publisher submission system might have a much larger return on investment in terms of automation. There is the risk that some web content pointed to by references will already have disappeared by this later editorial stage, but it is better that the author and editor can check this and together ensure that there is appropriate and sufficient evidence for what is stated in scholarly statement.

'Incorporating an archival step into a publisher submission system might have a much larger ROI'

The choice of platform for software engineering effort was again driven by the need for it to be free and open source, so the prototype was built using the Open Journal System (OJS).[17]

OJS plugin demonstrator

The plugin written for OJS creates a snapshot for each of the references in an article when it is uploaded to the system. Each time a new version of the article is uploaded, the references are archived to ensure that the author and/or editor are able to refer back to a snapshot of a reference at each stage of the submission process. It is also possible for the author to upload a list of archived URIs that was created manually or exported from another tool, such as Zotero. This shows the potential for the 'well-seamed' integration that can be achieved throughout the publication workflow in order to mitigate the risk of reference rot.

Stage 3: post publication

A third best time to act is at the post-publication stage. While not as preferable as Stage 1 or 2, obviously, post publication is better later than never as the reader needs assistance to locate the evidential feet on which scholarly statement stands. By the time the published work is on an access platform, being displayed on the reader's browser, the process of reference rot will have already begun. This is classic Memento[18] territory, in which the reader is assisted in travelling back in time to discover what of the web is held in web archives. The URI and indication of the date of citation (approximated by the date of publication, or better still the date of acceptance) is used to search across multiple web archives for versions of content pertaining to those URIs. Although not perfect, this is something that publisher

platforms can do to assist the reader, and of course the author whose work is otherwise insufficiently well published. Searching multiple archives for versions of URIs has been made easier with the Time Travel API[19] created by Herbert Van de Sompel and his team at Los Alamos National Laboratory Library.

Robust links: actionable temporal references

The addition of extra attributes to the HTML link, in order to indicate the location of the archived URI and the date and time at which the archive was made, creates a robust link.[20] This extra information is made available for software tools to use in order to direct readers of articles to the appropriate archived version of a reference, navigating across the scattered fragments of the preserved web.

It is important to note that replacing the URI with a potentially opaque URI to an archival service would only mean replacing one single point of failure with another one. The URI for the archived copy is provided in addition to the original URI, not as a replacement. Preserving the original URI enables the user to explore alternative routes to an archived version of the URI. This method of referencing archived URIs protects from failures in archiving services and gives the user the ability to follow their nose to find the appropriate archived version. The provision of the original URI and a date aids queries of other archival services to find an alternative version of the reference.

How to cite

Conventions for citing references to the web will need to change. An example of what needs to be included to make a link robust is shown in Figure 1.

'Conventions for citing references to the web will need to change'

Hiberlinks are modified <a> HTML elements that include the archive URL and timestamp as additional attributes:

```
<a href="http://www.newyorker.com/magazine/2015/01/26/cobweb"
data-versionurl="http://web.archive.org/web/20150219094636/http://
www.newyorker.com/magazine/2015/01/26/cobweb"
data-versiondate="2015-02-19T09:46:36">Cobweb Article</a>
```

Figure 1. Example of the robust link syntax for an article[21] published by *The New Yorker*

A version of this article showcasing the use of the robust link syntax is available on the Hiberlink website[22].

Conclusion: from Hyperlink to Hiberlink

The Hiberlink project has demonstrated that reference rot is a real and substantial threat to both the usability of what is published and to the integrity of the scholarly record. The project has defined and successfully raised awareness of the threat, which is now being given full attention in wider circles, as shown in recent coverage by *The New Yorker*[23]. Amusingly, this article is itself an example of a reference at the end of the URI that experienced content drift. The content and date changed from first moment of issue; the subtitle initially was: 'What the Web Said Yesterday'.

This article might have had the subtitle, 'From Hyperlink to Hiberlink', as that sums up both the challenge and the remedy proposed. What is required is an agreed set of methods by which content at the end of hyperlinks – content that authors regard as significant for their argument – is captured at an appropriate moment and held in a safe place ready to be awakened when the reader needs to check that reference.

This challenge of digital preservation extends beyond the content of articles in traditional serials to include ongoing 'integrating resources' such as databases and websites. The record of scholarship has a fuzzy edge. References are increasingly native to the web, interconnected and interlinked. This begs the question as to the extent to which data generated by the research process are themselves part of the record of scholarship[24]. It also involves considerations of what constitutes the copy (or copies) of records and notions of digital fixity.

'The record of scholarship has a fuzzy edge'

A start has been made. The Hiberlink project has also set about identifying and devising sustainable infrastructure that can support transactional archiving. This has been done in ways such that remedial action can be taken at various points in the life cycle of an article (and other forms of scholarly statement) by authors, editors, publishers, access platforms, archiving organizations and librarians.

However, it is just a start. Partners in the Hiberlink project are now reflecting on possible next steps. These include moving from prototype to production-quality tools to help remedy reference rot. It will also be necessary to anticipate issues associated with copyright and pay-wall access, which will obviously need careful consideration when proactively archiving.[25]

'possible next steps … include moving from prototype to production-quality tools to help remedy reference rot'

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
The authors would like to acknowledge the contributions of the Hiberlink project partners. In particular, the principal investigators: Claire Grover (for the Language Technology Group at the University of Edinburgh) and Herbert Van de Sompel (for the Los Alamos National Laboratory Library) and their research teams noted in the research articles referenced below.

References

1. The Hiberlink project: http://hiberlink.org/ (accessed 29 April 2015).
2. Zhou, K, Tobin, R and Grover, C, Extraction and Analysis of Referenced Web Links in Large-Scale Scholarly Articles, *Proceeding of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'14),* 2014, 451–452.
3. Klein, M, Van de Sompel, H, Sanderson, R, Shankar, H, Balakireva, L, Zhou, K and Tobin, R, Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLOS ONE*, 2014, 9(12): e115253; DOI: http://dx.doi.org/10.1371/journal.pone.0115253 (accessed 29 April 2015).
4. ETD2014: http://www2.le.ac.uk/library/etd2014 (accessed 30 April 2015).
5. Burnhill, P and Wincewicz, R, *Reference Rot: Threat and Remedy,* UKSG2015: http://www.slideshare.net/edinadocumentationofficer/reference-rot-and-linked-data-threat-and-remedy (accessed 30 April 2015).
6. Zhou K, et al., ref. 2.
7. Klein, M, et al., ref. 3.
8. Rosenthal, D, 10 February 2015, The Evanescent Web, DSHR's Blog: http://blog.dshr.org/2015/02/the-evanescent-web.html (accessed 9 June 2015).
9. Burnhill, P, Tales from The Keepers Registry: Serial Issues About Archiving & the Web, *Serials Review,* 39(1), 2013, 3–20; DOI: http://dx.doi.org/10.1016/j.serrev.2013.02.003 (accessed 30 April 2015).
10. Rusbridge, A, 4 May 2015, Preview our new service features, The Keepers Registry Blog: http://thekeepers.blogs.edina.ac.uk/ (accessed 30 April 2015).
11. Klein, M, et al., R, ref. 3.
12. Zittrain, J, Albert, K and Lessig, L, Perma: scoping and addressing the problem of link and reference rot in legal citations, *Harvard Law Review*, 2014, 127, 176–196.
13. Burnhill, P, *Reference Rot and E-Theses: Threat and Remedy:* http://www.slideshare.net/edinadocumentationofficer/reference-rotandetheses (accessed 29 April 2015).
14. Memento Time Travel: http://timetravel.mementoweb.org (accessed 30 April 2015).
15. Davis, R, Moving Targets: Web Preservation and Reference Management, *Ariadne*, January 2010, 62.
16. Zotero: https://www.zotero.org/ (accessed 30 April 2015).

17. PKP: Open Journal Systems:
https://pkp.sfu.ca/ojs/ (accessed 30 April 2015).

18. Van de Sompel, H, Nelson, M L and Sanderson, R, RFC 7089: HTTP Framework for Time-Based Access to Resource State – Memento, 2013.

19. Time Travels APIs:
http://timetravel.mementoweb.org/guide/api/ (accessed 30 April 2015).

20. Robust Links – Link Decoration:
http://robustlinks.mementoweb.org/spec/ (accessed 30 April 2015).

21. Lepore, J, 26 January 2015, The Cobweb – Can the Internet be archived? *The New Yorker*:
http://www.newyorker.com/magazine/2015/01/26/cobweb (accessed 30 April 2015).

22. Burnhill, P, Mewissen, M, Wincewicz, R, Reference rot in scholarly statement: threat and remedy, Insight, 2015.
http://hiberlink.org/Insight.html.

23. Lepore, J, ref. 21.

24. Burnhill, P, Mewissen, M and Rusbridge A, Where data and journal content collide: what does it mean to 'publish your data'? Edinburgh Research Archive:
https://www.era.lib.ed.ac.uk/handle/1842/9394 (accessed 30 April 2015).

25. Rosenthal, D, ref. 8.

Corresponding author: Muriel Mewissen

Project Manager
EDINA
The University of Edinburgh, Causewayside House, 160 Causewayside, Edinburgh EH9 1PR, UK
E-mail: muriel.mewissen@ed.ac.uk
ORCID ID: http://orcid.org/0000-0001-6809-5673

Peter Burnhill
ORCID ID: http://orcid.org/0000-0002-2654-7800

Richard Wincewicz
ORCID ID: http://orcid.org/0000-0002-3417-8865

To cite this article:
Burnhill, P, Mewissen, M and Wincewicz, R, Reference rot in scholarly statement: threat and remedy, *Insights*, 2015, 28(2), 55–61; DOI: http://dx.doi.org/10.1629/uksg.237

Published by UKSG in association with Ubiquity Press on 07 July 2015